MAGIC-CT: Multiorgan Annotation and Grounded Image Captioning in CT for Cancer

Maxim Popov

maxim.popov@alumni.nu.edu.kz

Nazarbayev University

Zangir Iklassov

Mohamed bin Zayed University of Artificial Intelligence

Zhanas Baimagambet

Nazarbayev University

Murat Jakipov

Nazarbayev University

Xeniya Andreyeva

Nazarbayev University

Muhammad Akhtar

Nazarbayev University

Martin Takáč

Mohamed bin Zayed University of Artificial Intelligence

Prashant Jamwal

Nazarbayev University

data-descriptor

Keywords: Computed Tomography, Abdominal Oncology, Multiorgan Segmentation, Grounded Image Captioning, Radiology Reports, 3D Medical Imaging

Posted Date: May 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8777425/v1

Additional Declarations: No competing interests reported.

MAGIC-CT: Multiorgan Annotation and Grounded Image Captioning in CT for Cancer

Maxim Popov[1,*], Zangir Iklassov[2], Zhanas Baimagambet[1], Murat Jakipov[3], Xeniya Andreyeva[3], Muhammad Akhtar[1], Martin Takáč[2], and Prashant Jamwal[1]

[1]Nazarbayev University, School of Engineering and Digital Sciences, Department of Computer Science, Astana, 010000, Kazakhstan
[2]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Department of Machine Learning, Abu Dhabi, United Arab Emirates
[3]National Research Oncology Center (NROC), Astana, Kazakhstan
[*]Correspondence: maxim.popov@alumni.nu.edu.kz
Author emails: maxim.popov@alumni.nu.edu.kz; zangir.iklassov@mbzuai.ac.ae; zhanas.baimagambet@nu.edu.kz; murat.jakipov@nu.edu.kz, nntsot@mail.ru; k.a.andreyeva@gmail.com; muhammad.akhtar@nu.edu.kz; prashant.jamwal@nu.edu.kz

ABSTRACT

Computed Tomography (CT) imaging is a cornerstone of abdominal oncology, offering critical insights into tumor morphology and spread. While artificial intelligence (AI) holds promise for automating lesion detection, segmentation, and reporting, progress is hindered by the scarcity of multimodal datasets that pair 3D anatomical annotations with expert-curated clinical descriptions. We present MAGIC-CT, a contrast-enhanced CT dataset of 562 patients with abdominal tumors (liver cysts/cancer, lung metastases, lung cancer, kidney cysts, renal cancer, pancreatic cancer). All 562 patients have CT scans and 3D lesion/organ masks; a subset of 492 patients also have organ-wise, radiologist-authored reports (RU/KZ/EN) totaling 4,937 organ descriptions. The dataset spans 8 pathologies across 4 organs, with about 1,250 annotated lesions and about 500 lesion-linked textual findings, enabling training of multimodal systems that connect volumetric localization to clinical language. MAGIC-CT uniquely integrates volumetric lesion localization, quantitative metrics (e.g., tumor volume, angular involvement of vasculature), and rich semantic context (e.g., "cuff-like encasement of the celiac trunk"), addressing the lack of resources bridging radiological imaging, segmentation, and clinical language. This dataset is expected to enable advancements in AI-driven tumor characterization, automated report generation, and metastasis tracking, with implications for precision oncology. This dataset is openly available at Zenodo: 10.5281/zenodo.18389015.

Keywords: Computed Tomography, Abdominal Oncology, Multiorgan Segmentation, Grounded Image Captioning, Radiology Reports, 3D Medical Imaging

Background & Summary

Accurate localization of abdominal tumors is nontrivial, given subtle imaging signatures and inter-organ variability. Deep within the body, organs like the liver, kidneys, and pancreas can harbor deadly malignancies that often grow silently. The scale of the challenge is staggering; hepatocellular carcinoma and pancreatic cancer alone claim over a million lives each year[1,2]. For many patients, the disease would have already spread by the time it's found, with nearly a third of new solid tumors having metastasized, most often to the liver[3]. This makes early and accurate detection not just a priority, but a race against time.

For decades, Computed Tomography (CT) has been our most powerful lens in this search, providing detailed cross-sectional maps of the body's internal landscape[4,5]. But this powerful tool places an immense burden on the human eye. Radiologists must meticulously scroll through hundreds of images, scrutinizing every slice for the subtlest signs of disease. They must distinguish benign cysts from malignant tumors, map a lesion's precise anatomical location, and hunt for ancillary clues like vascular invasion. This process is not only time-consuming but also fraught with challenges. The sheer volume of data, combined with intense time pressure and reader fatigue, can lead to diagnostic misses, especially for small or low-contrast lesions. While structured reporting systems exist to standardize this process, their adoption is inconsistent[6], leaving the radiologist's skill and vigilance as the last line of defense.

In this context, deep learning has shown considerable promise. An AI model can process an entire CT scan in seconds, acting as a tireless second reader that never fatigues. It can highlight suspicious areas, automatically segment tumors, and track their size over time. Recent breakthroughs have demonstrated a tangible benefit. In one major trial, an AI model's ability

to segment liver metastases was on par with or better than that of radiology fellows, significantly improving the accuracy of treatment response assessment[7]. In another landmark study, a DL system not only matched radiologists in detecting pancreatic cancer but also identified four tumors that human readers had initially missed, some measuring less than two centimeters[8]. By augmenting radiologist expertise, AI can accelerate detection and improve reliability.

Significant progress in medical artificial intelligence (AI) is held back by a fundamental bottleneck: a lack of comprehensive, publicly available datasets. To be truly useful in a clinical setting, an AI model needs to learn from data that mirrors a radiologist's workflow, which involves both visual interpretation (the pixels) and diagnostic reasoning (the prose of a report). However, existing resources are fragmented, typically offering one without the other. On one hand, we have datasets rich in pixel-level annotations. Many of these are narrowly focused on a single organ, such as the Liver Tumor Segmentation (LiTS) dataset for the liver[9] or the Kidney Tumor Segmentation 2023 (KiTS23) dataset for the kidneys[10]. This limited scope hinders the development of models that can analyze the complex, multi-organ landscape of abdominal oncology. The consequences are clear: models trained on these datasets perform well at segmenting the whole organ but struggle significantly with the more critical task of identifying tumors. For example, in the LiTS challenge, the best tumor segmentation accuracy was over 20 points lower than for whole-liver segmentation and degraded further when tested at different medical centers[9]. While larger datasets like AbdomenCT-1K—an abdominal computed tomography (CT) collection with multi-organ labels[11]—offer robust multi-organ segmentation, they lack the corresponding radiological reports, leaving the model without crucial diagnostic context.

Recognizing the above detailed gap, a few recent efforts have tried to unite pixels and prose. RadGenome-Chest CT stands out as a landmark achievement in thoracic imaging, providing over 25,000 CT volumes with both organ-level masks and detailed textual descriptions[12]. It serves as a blueprint for what is needed, but its focus on the chest leaves a critical void in abdominal imaging. An attempt to fill this void is AbdomenAtlas 3.0[13], which pairs over 9,000 abdominal CT scans with detailed tumor segmentations. However, instead of using expert-written reports, it relies on AI-generated synthetic text. This approach carries significant risk, as synthetic reports can introduce subtle inaccuracies or outright "hallucinations" that are unacceptable for clinical applications. They also tend to focus only on the organs with pathology, neglecting the important descriptions of surrounding healthy structures. Thus, a critical gap remains: there is no large-scale, public dataset for abdominal imaging that combines precise, expert-validated segmentation masks with the rich, nuanced diagnostic information of human-authored radiology reports. To address this critical need, we introduce MAGIC-CT: a new dataset designed to bridge the gap between pixels and prose in abdominal oncology. This resource is the first to pair high-resolution, 3D segmentation masks with the corresponding expert-written radiology reports for 562 patients with diverse oncologic conditions. The segmentations delineate eight key abdominal organs, including the liver, lungs, kidneys, and pancreas—along with any associated pathologies. The paired narrative reports provide the crucial clinical context, offering a comprehensive assessment of all visible structures and characterizing abnormalities by their size, location, and morphology. By unifying these two data streams, MAGIC-CT provides the foundation for a new class of AI models: those that can simultaneously segment anatomy and generate meaningful diagnostic text. The goal is to create tools that function more like a human expert, not only highlighting a tumor but also describing its clinical significance. This integrated approach promises to enhance the accuracy and interpretability of AI-driven diagnostics, paving the way for more reliable tools in the fight against abdominal cancer. While MAGIC-CT centers on abdominal oncology, thoracic involvement relevant to abdominal primaries (e.g., lung metastases) is included to support end-to-end oncologic workflows.

Paper organization. The rest of the paper is organized as follows. In **Methods**, we describe the cohort, imaging protocols, annotation workflow, and ethical approvals. **Data Records** details the public release (file formats, folder structure, and tri-lingual organ-wise reports). **Technical Validation** presents inter-annotator agreement and baseline segmentation results across multiple architectures with size- and organ-stratified analyses. **Usage Notes** outlines intended applications and known limitations. **Code Availability** points to loaders, preprocessing, and example scripts. Finally, we provide **Acknowledgements**, **Author Contributions**, and **Competing Interests**.

Methods

Ethical Statement

This retrospective study using fully de-identified clinical imaging data was approved by the *Nazarbayev University Institutional Research Ethics Committee (NU-IREC)* via expedited review, protocol 1099/08072025, approved on 27 August 2025 and valid through 26 August 2026. Secondary use and data sharing were additionally approved by the *National Research Oncology Center (NROC) Ethics Committee*, approval No. 34/30.07.2025. All imaging data were de-identified at NROC before transfer; the re-identification key remains solely at NROC. NROC (data controller) and Nazarbayev University (data recipient/processor) executed a Data Use/Transfer Agreement governing secure transfer, storage, and permitted uses. Given the retrospective design and full de-identification, both committees waived the requirement for informed consent.

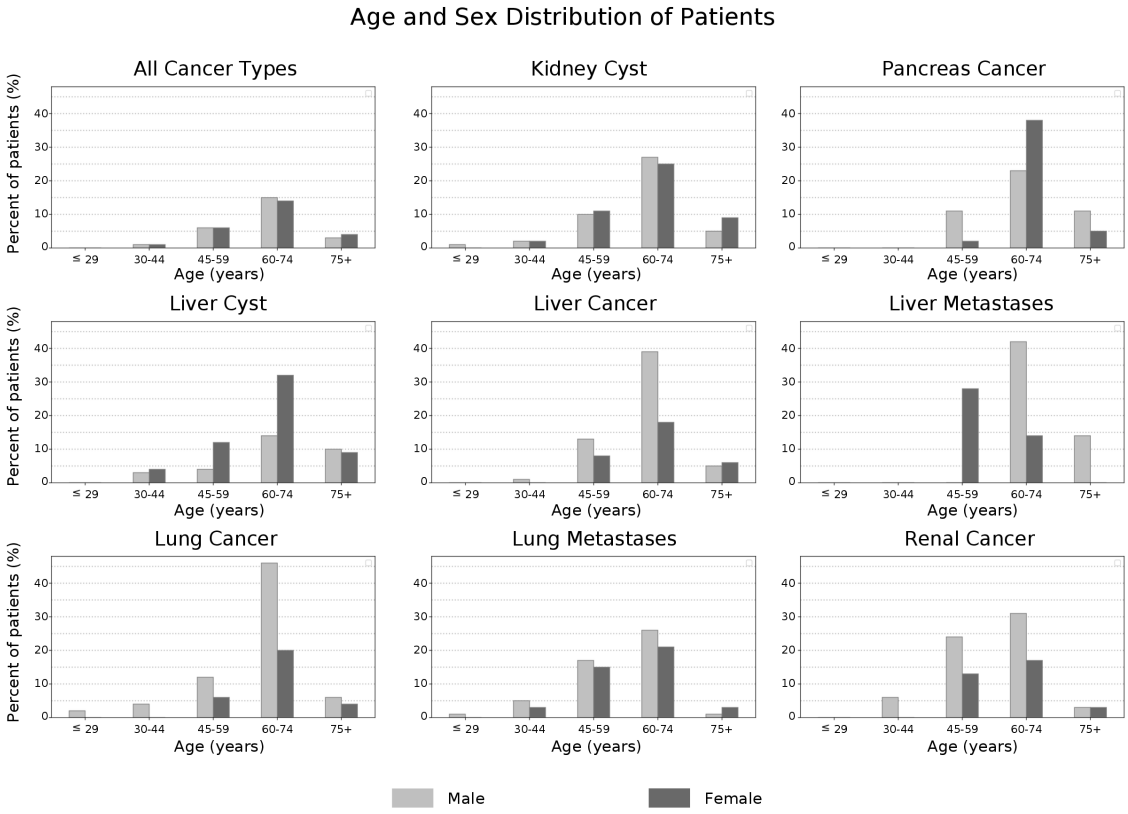


Figure 1. Age and Gender distribution of patients across diagnostic objectives.

Patient Cohort

The study cohort consists of patients with confirmed cases, whose clinical data are available at the National Research Oncology Center, Astana, Kazakhstan. The total number of unique patients is 562, with a mean age of 63 years and a median age of 64; 52% are men (youngest 21, oldest 92) and 46% are women (youngest 19, oldest 91). Of the 562 patients with imaging and segmentations, 492 include tri-lingual organ-wise reports. The 70 patients without reports remain available for segmentation-only benchmarks. Full demographic breakdown for different imaging objectives is shown in Figure 1.

The selection of subjects and scans for this study was guided by a combination of clinical and technical inclusion criteria to ensure diagnostic relevance and high-quality annotations. Clinically, only cases with a confirmed diagnosis—validated through histopathological findings, formal physician reports, or discharge summaries were included. To capture the diversity of abdominal oncology, the dataset encompasses a wide range of tumor types and anatomical locations, including both benign and malignant lesions in the liver, lungs, kidneys, and pancreas. Particular emphasis was placed on selecting tumors of medium to large size, as these are more readily visualized and accurately segmented. Cases exhibiting significant artifacts or imaging complications, such as ascites, widespread metastases, or metallic implants, were excluded to maintain clarity and consistency in the data.

Imaging

The imaging data for the MAGIC-CT dataset were acquired using a Philips Ingenuity (Philips Healthcare, Eindhoven, The Netherlands) CT scanner, ensuring consistency and high-quality imaging across all included patients. The scans were performed with intravenous administration of contrast agents—primarily Ultravist 370, an iodine-based contrast medium optimized for enhanced soft-tissue differentiation, and Gadovist in selected cases, depending on clinical indications. These contrast agents facilitated improved visualization of vascular structures, organ parenchyma, and lesion margins.

Radiation exposure was carefully managed, with scans performed at an average effective radiation dose ranging from 7 to 15 millisieverts (mSv). This range reflects standard clinical practice, balancing image quality with radiation safety considerations. Such consistent imaging protocols ensured uniform data quality, which is crucial for accurate segmentation, lesion characterization, and subsequent integration into AI-based diagnostic modeling.

On the technical side, only scans with full anatomical coverage of the target organ across all CT slices were considered

Table 1. Summary statistics of segmented oncology datasets in CT scans. Values are reported as mean ± standard deviation.

Dataset (Organ/Cancer)	Num Segmented Components	Segmentation Volume (mm^3)	Scan Volume (mm^3)
Liver Cancer	1.49 ± 1.93	538.30 ± 723.81	60608.17 ± 29523.17
Renal Cancer	1.14 ± 0.64	98.85 ± 153.69	72612.57 ± 31081.56
Liver Cyst	0.72 ± 1.18	28.27 ± 90.56	68209.12 ± 27996.26
Pancreas	1.09 ± 0.29	36.47 ± 33.43	49164.65 ± 24733.85
Kidney Cyst	0.99 ± 1.58	343.48 ± 4895.73	74742.35 ± 59865.20
Lung Metastases	1.24 ± 2.05	7.99 ± 16.32	59110.87 ± 21214.07
Lung Cancer	1.06 ± 0.37	176.39 ± 392.80	72165.92 ± 26877.96
Liver Metastases	3.00 ± 4.97	17.55 ± 32.72	85476.89 ± 42963.57

suitable for annotation. For liver lesions, contrast-enhanced studies were mandatory to ensure adequate differentiation between lesions and the surrounding parenchyma. In general, selected scans were required to be free from major imaging artifacts, including those caused by patient motion, metallic objects, or intestinal peristalsis. This dual-layered selection process ensured that each case included in the dataset met both clinical validity and technical quality standards necessary for effective model training and evaluation.

Table 1 summarizes key statistics from a collection of segmented oncology datasets in CT scans, encompassing a range of organs and lesion types. For each dataset, the average number of segmented components, segmentation volume (mm^3), and scan dimension (voxel-based image size) are reported as mean ± standard deviation. These statistics provide an overview of the dataset diversity in terms of lesion burden, anatomical scale, and variability, supporting downstream analyses and method benchmarking. This dataset demonstrates substantial heterogeneity in segmented oncology CT scans, with tumor volumes ranging from small metastatic lesions (mean $\approx 8mm^3$ for lung metastases) to much larger primary tumors (mean$\approx 538mm^3$ for hepatocellular carcinoma). The number of segmented components also varies, with primary tumors typically presenting as single entities, whereas metastatic and cystic lesions show more fragmented or multiple regions. Notably, high coefficients of variation in both volume and component count across most categories highlight the diversity and variability of tumor burden and morphology within and between organ systems. These findings underscore the importance of organ-specific analysis pipelines and standardized segmentation protocols for reliable downstream study and benchmarking.

Annotation

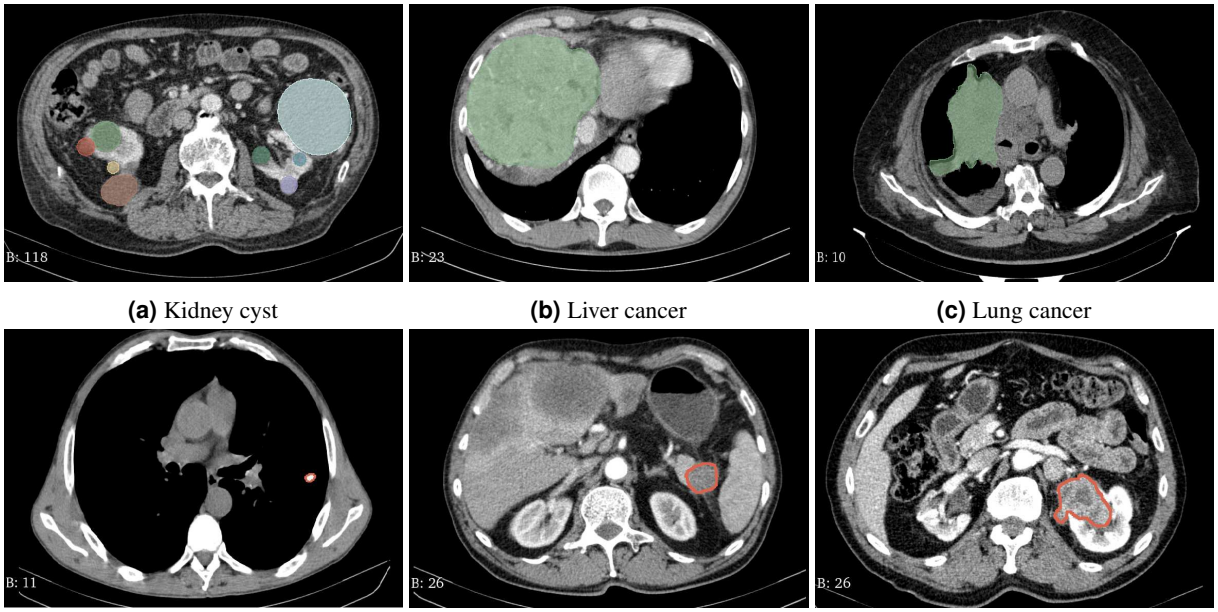
Annotations for the MAGIC-CT dataset were systematically created using the open-source medical imaging software, 3D Slicer. Radiologists used the software's comprehensive segmentation tools to generate precise three-dimensional masks of primary organs and lesions visible in abdominal CT scans. Each annotator underwent specialized training to standardize segmentation protocols and ensure consistency across the dataset.

The segmentation workflow involved loading CT volumes into 3D Slicer, followed by manual delineation of anatomical structures and pathological lesions. To enhance precision, annotators utilized semi-automated segmentation tools within the software, such as threshold-based segmentation, region-growing methods, and boundary refinement algorithms. Annotators carefully inspected and adjusted each mask slice-by-slice in axial, coronal, and sagittal planes, ensuring accurate volumetric delineations. Particular attention was paid to capturing all clinically relevant lesions. Lesions were annotated based on clear radiological criteria, including distinct margins, density differences from surrounding tissues, and known tumor characteristics. Following initial segmentation, a senior radiologist reviewed each annotation to verify accuracy and completeness, providing feedback or corrections as necessary. Through this rigorous annotation approach utilizing 3D Slicer, MAGIC-CT achieves high-quality, consistent segmentation masks, ensuring robust data suitable for developing advanced diagnostic models in abdominal oncology. The examples of annotations across organs are shown in Figure 2.

The captioning strategy employed across all patient cases involves a detailed, narrative-style approach that comprehensively describes the condition of all visible abdominal organs. Each narrative report meticulously captures essential clinical insights, including organ-specific anatomical descriptions, precise lesion measurements, and the characterization of abnormalities. The organs routinely assessed in these reports encompass the liver, pancreas, gallbladder, bile ducts, spleen, kidneys, adrenal glands, stomach, duodenum, lymph nodes, and relevant bony structures.

Narrative captions systematically include quantitative measurements, particularly lesion dimensions, and qualitatively describe the nature of abnormalities, such as tumors, cysts, infiltrative lesions, structural anomalies, or fluid accumulations. This thorough documentation approach ensures that each generated report accurately mirrors clinical radiological assessment practices, providing rich context that supports meaningful integration with segmentation masks, thereby enhancing the development and evaluation of advanced diagnostic models.

Table 2 highlights broad coverage (10.0 ± 2.6 organs per patient; 17.8 ± 19.1 tokens per description) and variability in



(a) Kidney cyst **(b)** Liver cancer **(c)** Lung cancer

(d) Metastatic lung view **(e)** Pancreatic cancer **(f)** Renal cancer

Figure 2. Axial CT views showing various organs and pathologies from the MAGIC-CT dataset.

Table 2. Quantitative details of the organ captioning dataset, covering the number of organ descriptions, patients, average number of organs described per patient (mean ± std), and average number of tokens per description (mean ± std).

Dataset	Descriptions	Subjects	Organs/Patient	Tokens/Description
HCC	461	45	10.2 ± 2.0	20.0 ± 23.5
Kidney Cyst	2082	202	10.3 ± 2.5	18.8 ± 19.8
Lung MTS	385	41	9.4 ± 3.1	15.7 ± 12.5
Pancreas	303	29	10.4 ± 2.8	15.5 ± 15.4
Renal Cancer	309	29	10.7 ± 1.4	17.0 ± 19.0
Liver MTS	63	7	9.0 ± 1.2	17.3 ± 17.1
Liver Cyst	976	96	10.2 ± 2.4	15.9 ± 17.1
Lung Cancer	358	43	8.3 ± 3.5	19.1 ± 21.3
TOTAL	4937	492	10.0 ± 2.6	17.8 ± 19.1

reporting. The pathology distribution is imbalanced—kidney cysts constitute the largest subset (2,082 descriptions across 202 patients), whereas liver metastases are the smallest (63 descriptions across 7 patients)—indicating a long-tail regime. We therefore report per-class metrics (Dice/HD95, sensitivity/specificity) alongside macro-averages, and recommend class-balanced sampling or loss reweighting when training downstream models.

Data Records

The dataset is archived and citable via Zenodo: 10.5281/zenodo.18389015 (version DOI; latest version via the Concept DOI 10.5281/zenodo.17549292).

The dataset structure is as follows:

All imaging data in the MAGIC-CT dataset are stored in a separate "scans" folder in anonymized .nrrd format, which preserves essential volumetric imaging information, such as pixel spacing, slice thickness, and spatial orientation, while ensuring complete removal of any personally identifiable information. This format supports efficient loading into medical image analysis platforms and maintains compatibility with major deep learning libraries for 3D processing. Segmentation masks are also stored in .nrrd format in a folder named "segmentations", and follow the same anonymization and spatial alignment conventions. Each mask is registered to the corresponding scan and labeled according to the annotated organ or pathology.

Textual reports and associated metadata are organized by pathology and stored in structured .json files, ensuring clarity,

modularity, and ease of integration. Each .json file contains key-value fields capturing basic patient metadata (such as age, gender, and scan date) and three language-specific blocks: "ru" (Russian), "kz" (Kazakh), and "en" (English). Within each language block, reports are organized by anatomical structure, allowing direct access to narrative findings per organ.

Visually, the structure of the reports is as follows:

```
{
  "patient_id": "...",
  "age": 65,
  "gender": "male",
  "en": {
    "liver": "The liver has a homogeneous structure; no focal lesions detected.",
    "pancreas": "Mass identified in the pancreatic body, measuring $5.3x3.5$ cm...",
    "kidneys": "Small cortical cysts in both kidneys, no hydronephrosis.",
    ...
  }
}
```

Evaluation metrics. We report the Dice similarity coefficient (DSC), the 95th-percentile Hausdorff distance (HD95), sensitivity, and specificity, along with runtime/resource indicators. For two binary masks A (prediction) and B (reference), with true positives (TP), false positives (FP), and false negatives (FN):

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}. \tag{1}$$

The Hausdorff distance $H(A,B)$ is the maximum surface-to-surface distance; HD95 is its 95th percentile to reduce outlier effects. Voxel-wise classification metrics are

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{2}$$

Operational measures include *Inference Time (s)*: wall-clock seconds per volume for a single forward pass; *Parameters (M)*: trainable parameter count in millions; and *Memory (GB)*: peak GPU memory during inference with the stated input size.

Technical Validation

Annotation assessment

To evaluate the consistency of image labeling among the seven participating radiologists, we measured their inter-annotator agreement using Cohen's kappa (κ). This robust statistical metric is ideal for this purpose as it quantifies the degree of consensus between raters while correcting for agreement that could occur by chance. To accommodate the seven raters, pairwise Kappa scores were computed for all combinations of radiologists. The Kappa coefficient is calculated using the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3}$$

where p_o (Observed Agreement) is the proportion of items where the two raters actually agreed. It is calculated by dividing the number of items they agreed on by the total number of items, and p_e (Expected Chance Agreement) is the probability that the raters would agree by chance. It is calculated by considering the individual rating patterns of each rater. For each category we multiply the proportion of items one rater assigned to that category by the proportion the other rater assigned to it. The final p_e is the sum of these products across all categories. The resulting Kappa score is interpreted on a scale from -1 to 1. A score of 1 signifies perfect agreement, 0 indicates that the observed consensus is no better than random chance, and a negative score implies systematic disagreement. Higher positive values correspond to a stronger level of agreement. For qualitative assessment, these scores are often categorized using standard benchmarks, such as the Landis and Koch scale, which defines levels of agreement from "Slight" to "Almost Perfect."

Table 3 summarizes inter-annotator agreement across patients and readers. The overall mean Cohen's kappa was 0.74, indicating "Substantial" agreement among radiologists and demonstrating reliability well above chance.

Table 3. Average inter-annotator agreement across patients, measured using Cohen's Kappa score.

		Patients										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**
Doctors	1	0.96	0.63	0.70	0.52	0.93	0.83	0.61	0.79	0.06	0.95	0.70
	2	0.95	0.57	0.76	0.43	0.93	0.73	0.55	0.60	0.55	0.95	0.70
	3	0.94	0.49	0.66	0.48	0.90	0.82	0.58	0.77	0.64	0.95	0.73
	4	0.96	0.59	0.82	0.58	0.95	0.87	0.64	0.81	0.77	0.96	0.80
	5	0.95	0.60	0.69	0.55	0.94	0.87	0.62	0.71	0.71	0.95	0.71
	6	0.95	0.60	0.78	0.46	0.94	0.86	0.62	0.79	0.66	0.96	0.76
	7	0.96	0.59	0.82	0.58	0.95	0.87	0.64	0.81	0.77	0.96	0.80
	Mean	0.95	0.58	0.72	0.53	0.93	0.84	0.61	0.75	0.60	0.96	0.74

Table 4. Overall segmentation performance of pretrained models on MAGIC-CT (inference-only).

Model	**Overall Dice (%)**	**Inference Time (s)**	**Parameters (M)**	**Memory (GB)**
SwinUNETR	**72.3 ± 2.1**	2.8	61.9	10.2
UNETR	68.7 ± 2.3	2.5	102.4	12.8
SegResNet	65.2 ± 1.9	**1.2**	**15.7**	**6.4**
DynUNet	67.1 ± 2.0	1.8	22.3	7.8

Deep Learning model training

Baseline Models. We evaluated four representative deep learning architectures spanning the evolution of medical image segmentation: two transformer-based models—SwinUNETR[14] and UNETR[15]—and two convolutional neural network (CNN) approaches—SegResNet[16] and DynUNet (configured following nnU-Net principles)[17]. Implementations and pretrained weights were used via the MONAI framework[18]. SwinUNETR combines a Swin Transformer encoder with a U-Net–style decoder for precise 3D boundary delineation[14], while UNETR pioneered the use of Vision Transformers for 3D medical image segmentation by leveraging global self-attention to capture long-range dependencies[15]. SegResNet serves as a strong CNN baseline derived from residual encoder–decoder designs with autoencoder regularization shown to generalize well across medical tasks[16]. DynUNet adopts an automatically configured topology (patch size, kernel sizes, and depth) guided by dataset properties, embodying the self-configuring strategy established by nnU-Net[17]. Together, these models cover established CNN paradigms and emerging transformer architectures, enabling a rigorous evaluation on the proposed multi-cancer segmentation benchmark.

Model Training and Evaluation Protocol. To ensure reproducible baselines and eliminate training-induced variability, we employed an inference-only evaluation protocol using established pre-trained models from the MONAI framework[18]. SwinUNETR utilized self-supervised pre-trained encoder weights derived from 5,050 publicly available CT scans across multiple anatomical regions, trained using reconstruction, rotation prediction, and contrastive learning objectives[19]. Pre-trained weights for all models were obtained from validated open-source repositories, with SwinUNETR achieving state-of-the-art performance on standardized benchmarks[14]. All input volumes were standardized to 96^3 voxels with consistent preprocessing pipelines. Evaluation metrics comprised Dice similarity coefficient, 95th percentile Hausdorff Distance (HD95), sensitivity, and specificity, calculated per cancer type and aggregated using equal weighting to prevent dataset size bias[20]. Statistical significance was assessed using paired t-tests, and effect sizes were quantified using Cohen's d to establish clinical relevance thresholds. This methodology enables direct architectural comparison while leveraging validated pre-training from the medical imaging community.

Table 4 reports that transformer-based architectures lead multi-cancer CT segmentation on MAGIC-CT. Dice scores were averaged with equal weighting across seven cancer types (hepatocellular carcinoma, renal cancer, liver cysts, pancreatic cancer, kidney cysts, lung metastases, and lung cancer) to avoid size-driven bias. SwinUNETR attained the highest overall Dice ($72.3 \pm 2.1\%$), exceeding UNETR ($68.7 \pm 2.3\%$) by 3.6 points and SegResNet ($65.2 \pm 1.9\%$) by 7.1 points, underscoring the benefit of attention-based encoders. DynUNet ($67.1 \pm 2.0\%$) outperformed the traditional SegResNet baseline, indicating gains from self-configuring U-Net designs. Inference times of 1.2–2.8 s per volume further support the feasibility of all evaluated models for near–real-time use.

Table 5 provides a cancer-type–specific evaluation that highlights the heterogeneity inherent in multi-cancer segmentation benchmarks. Performance was computed independently for each cancer type using patient-specific ground truth annotations,

Table 5. Cancer-type Dice (%) on MAGIC-CT.

Cancer Type	Model Performance (Dice %)			
	SwinUNETR[14]	UNETR[15]	SegResNet[16]	DynUNet (nnU-Net)[17]
Benign Lesions				
Liver Cysts	**84.3 ± 3.2**	80.7 ± 3.8	76.9 ± 4.1	79.3 ± 4.5
Kidney Cysts	**81.0 ± 3.8**	77.6 ± 4.2	73.6 ± 4.6	76.0 ± 4.8
Primary Malignancies				
Hepatocellular Carcinoma	**78.1 ± 4.2**	74.5 ± 4.6	70.5 ± 5.1	72.8 ± 5.3
Lung Cancer	**74.9 ± 4.8**	71.4 ± 5.2	67.7 ± 5.5	69.6 ± 5.7
Renal Cancer	**71.6 ± 5.1**	68.0 ± 5.4	64.4 ± 5.8	66.2 ± 6.0
Pancreas Cancer	**53.1 ± 7.8**	49.3 ± 8.1	46.8 ± 8.4	47.9 ± 8.6
Metastatic Disease				
Lung Metastases	**63.0 ± 6.2**	59.4 ± 6.8	56.5 ± 7.1	57.9 ± 7.4
Average	**72.3 ± 2.1**	68.7 ± 2.3	65.2 ± 1.9	67.1 ± 2.0

Table 6. Organ-wise segmentation with SwinUNETR on MAGIC-CT: Dice, sensitivity, specificity, and HD95.

Organ System	Dice (%)	Sensitivity (%)	Specificity (%)	HD95 (mm)
Large Organs				
Liver	**81.2 ± 4.1**	**82.8 ± 3.8**	98.7 ± 0.8	8.2 ± 2.1
Lungs	71.8 ± 4.9	73.4 ± 4.6	**99.2 ± 0.6**	**6.8 ± 1.9**
Medium Organs				
Kidneys	74.3 ± 5.2	75.9 ± 5.1	98.9 ± 0.7	9.4 ± 2.8
Small Organs				
Pancreas	53.1 ± 7.8	58.5 ± 8.2	99.4 ± 0.5	18.3 ± 5.7

revealing a clear hierarchy of difficulty. Benign lesions achieved the highest accuracy—with liver cysts at $84.3 \pm 3.2\%$ and kidney cysts at $81.0 \pm 3.8\%$ using SwinUNETR—consistent with their well-defined boundaries and homogeneous appearance. Primary malignancies showed intermediate performance, ranging from $78.1 \pm 4.2\%$ for hepatocellular carcinoma to $53.1 \pm 7.8\%$ for pancreatic cancer, the latter being especially challenging due to low contrast and irregular tumor margins. Metastatic disease was similarly difficult, with lung metastases at $63.0 \pm 6.2\%$, reflecting variability in lesion morphology and distribution. Across all cancer types, SwinUNETR maintains an 8–15 percentage-point advantage over other baselines, supporting the robustness of transformer-based encoders for diverse oncologic segmentation tasks and establishing this benchmark as a comprehensive testbed for automated cancer detection.

Table 6 examines organ-specific performance with SwinUNETR, linking anatomical complexity to segmentation accuracy. Results aggregate performance across all pathological conditions within each organ, highlighting organ-specific challenges. Large organs achieved the highest accuracy, with liver segmentation reaching 81.2 (±) 4.1% Dice and 82.8 (±) 3.8% sensitivity, benefiting from substantial volume and relatively consistent imaging characteristics. Lung segmentation attained 71.8 (±) 4.9% Dice and the highest specificity at 99.2 (±) 0.6%, reflecting accurate background rejection (i.e., a low false-positive rate) despite the difficulty of detecting both primary and metastatic lesions. Medium-sized organs, represented by kidneys at 74.3 (±) 5.2% Dice, showed intermediate performance consistent with moderate anatomical complexity. Pancreatic segmentation was the most challenging at 53.1 (±) 7.8% Dice, with the highest HD95 of 18.3 (±) 5.7 mm, where HD95 denotes the 95th-percentile Hausdorff distance between predicted and reference surfaces; this is consistent with known issues of small organ size, low contrast, and irregular tumor boundaries. Consistently high specificity (98.7–99.4%) across organs indicates reliable separation of pathological regions from healthy tissue in this benchmark.

Table 7 stratifies performance by tumor size and lesion multiplicity, revealing key factors that influence automated segmentation across all architectures. Tumor size is defined as the *maximum axial diameter on CT slices*, and performance was computed by grouping lesions accordingly and by multiplicity from radiological annotations. Large tumors (>5 cm) achieved the best accuracy with SwinUNETR at $76.8 \pm 4.2\%$, medium tumors (2–5 cm) dropped to $65.4 \pm 5.8\%$, and small tumors (<2 cm) were most challenging at $51.2 \pm 8.1\%$, illustrating the expected decline in performance with decreasing lesion size.

Table 7. Segmentation by tumor size and lesion multiplicity across pretrained models.

Tumor Characteristic	SwinUNETR	UNETR	SegResNet	DynUNet
Tumor Size				
Large (>5 cm)	**76.8 ± 4.2**	74.1 ± 4.6	71.9 ± 5.1	70.3 ± 5.4
Medium (2-5 cm)	**65.4 ± 5.8**	62.7 ± 6.2	60.8 ± 6.5	59.1 ± 6.8
Small (<2 cm)	**51.2 ± 8.1**	47.8 ± 8.6	45.3 ± 9.1	43.7 ± 9.4
Multiplicity				
Single Lesion	**72.1 ± 4.8**	69.4 ± 5.2	67.3 ± 5.6	65.8 ± 5.9
Multiple Lesions	**58.9 ± 7.2**	55.8 ± 7.7	53.6 ± 8.1	52.1 ± 8.4

Table 8. Paired model comparisons: p-values and effect sizes (Cohen's d).

Model Comparison	Dice Score		HD95 Distance		Clinical Significance
	p-value	Cohen's d	p-value	Cohen's d	
SwinUNETR vs UNETR	0.003**	0.52	0.008**	-0.48	Moderate
SwinUNETR vs SegResNet	<0.001***	0.89	<0.001***	-0.82	Large
SwinUNETR vs DynUNet	<0.001***	0.65	<0.001***	-0.61	Large
UNETR vs SegResNet	0.041*	0.44	0.052	-0.41	Small
UNETR vs DynUNet	0.028*	0.18	0.035*	-0.15	Small
Transformers vs CNNs	<0.001***	1.12	<0.001***	-1.05	Very Large

Multiplicity analysis shows single-lesion cases at $72.1 \pm 4.8\%$ versus $58.9 \pm 7.2\%$ for multi-lesion cases—a 13.2 percentage-point degradation reflecting the complexity of multi-focal disease; *multiple lesions are primarily observed in lung metastases and other multi-focal tumors*. The consistent performance ordering (SwinUNETR > UNETR > DynUNet > SegResNet) across strata supports the robustness of these findings and establishes tumor size and multiplicity as core benchmarking criteria for oncologic segmentation.

Table 8 presents paired significance tests that confirm the clinical relevance of the observed performance gaps. P-values were computed using paired t-tests across cancer types ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$), and practical significance was quantified with Cohen's d (interpreted as 0.2 small, 0.5 medium, 0.8 large, 1.2 very large). The largest effect size occurs for SwinUNETR versus SegResNet ($d = 0.89$, $p < 0.001$), indicating a "large" improvement beyond typical diagnostic thresholds. SwinUNETR is statistically superior to all other architectures, with effect sizes spanning 0.52–0.89, and the aggregate comparison of transformer- versus CNN-based models yields a "very large" effect ($d = 1.12$, $p < 0.001$). For HD95, negative d values indicate better performance (lower boundary distance), mirroring the Dice gains and signaling clinically meaningful contour accuracy. Overall, these tests are based on inference results aggregated across cancer types and show that the benchmark reliably resolves clinically significant differences between segmentation algorithms.

Usage Notes

MAGIC-CT is archived at Zenodo: 10.5281/zenodo.18389015. The dataset is released under the CC0-1.0 license. To access the dataset, no registration or other controls are required.

Image processing techniques, such as contrast enhancement and the binary segmentation model, as well as annotation conversion tools used in this article, can be found in the repository under the Code Availability section. All related project data are freely available under a CC0 license.

Code Availability

Scripts used to process the dataset, as well as to compute the inter-annotator agreement, are located in project GitHub. Scripts related to segmentation model training and evaluation are located in a separate repositoryhttps://github.com/Zangir/MAGIC-CT/tree/main

Funding

This work was supported by the Collaborative Research Program of Nazarbayev University (Grant No. 111024CRP2007).

Acknowledgements

The authors acknowledge the Research Office of the National Research Oncology Center (NROC), Astana, Kazakhstan, for institutional support and coordination of this research, and are grateful to the physicians and medical staff at NROC for their clinical expertise and valuable contributions to data processing and annotation.

Author contributions statement

P. Jamwal[1] served as the team lead and the facilitator of the research project. M. Popov organized and managed the data, developed data processing scripts, computed inter-annotator agreement, compiled related works, and wrote the manuscript. Z. Iklassov prepared the baseline assessment section and trained the baseline models. Z. Baimagambet[1], M. Jakipov[3], X. Andreyeva[3], and the National Research Oncology Center (NROC) annotation team performed all data annotation and facilitated the Patient Cohort and Imaging section writing process. M. Takáč served as a final editor and assisted in writing and proofreading the manuscript. All authors reviewed the manuscript, provided critical feedback, and approved the final version.

Competing interests

The authors declare no competing interests.

References

1. Rumgay, H. *et al.* Global burden of primary liver cancer in 2020 and predictions to 2040. *J. Hepatol.* **77**, 1598–1606, 10.1016/j.jhep.2022.08.021 (2022).
2. Ushio, J. *et al.* Pancreatic ductal adenocarcinoma: Epidemiology and risk factors. *Diagnostics* **11**, 562, 10.3390/diagnostics11030562 (2021).
3. Jamil, A. & Kasi, A. Lung metastasis. In *StatPearls [Internet]* (StatPearls Publishing, 2023).
4. Chezmar, J. *et al.* Liver and abdominal screening in patients with cancer: Ct versus mr imaging. *Radiology* **168**, 43–47 (1988).
5. Caraiani, C., Yi, D., Petresc, B. & Dietrich, C. Indications for abdominal imaging: When and what to choose? *J. ultrasonography* **20**, e43 (2020).
6. Elsayes, K. M. *et al.* Li-rads: a conceptual and historical review from its beginning to its recent integration into aasld clinical practice guidance. *J. hepatocellular carcinoma* 49–69 (2019).
7. Liu, X. *et al.* Automatic segmentation of hepatic metastases on dwi images based on a deep learning method: assessment of tumor treatment response according to the recist 1.1 criteria. *BMC cancer* **22**, 1285 (2022).
8. Park, H. J. *et al.* Deep learning–based detection of solid and cystic pancreatic neoplasms at contrast-enhanced ct. *Radiology* **306**, 140–149 (2023).
9. Bilic, P. *et al.* The liver tumor segmentation benchmark (lits). *Med. Image Analysis* **84**, 102680, 10.1016/j.media.2022.102680 (2023).
10. Heller, N. *et al.* The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct (2023). 2307.01984.
11. Ma, J. *et al.* Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 6695–6714, 10.1109/TPAMI.2021.3100536 (2022).
12. Zhang, X. *et al.* Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis (2024). 2404.16754.
13. Bassi, P. R. *et al.* Radgpt: Constructing 3d image-text tumor datasets. *arXiv preprint arXiv:2501.04678* (2025).
14. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266* (2022).
15. Hatamizadeh, A. *et al.* Unetr: Transformers for 3d medical image segmentation. *Proc. IEEE/CVF winter conference on applications computer vision* 574–584 (2022).
16. Myronenko, A. 3d mri brain tumor segmentation using autoencoder regularization. *Int. MICCAI Brainlesion Work.* 311–320 (2019).
17. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. methods* **18**, 203–211 (2021).

18. Cardoso, M. J. *et al.* Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701* (2022).

19. Tang, Y. *et al.* Self-supervised pre-training of swin transformers for 3d medical image analysis. *Proc. IEEE/CVF Conf. on Comput. Vis. Pattern Recognit.* 20730–20740 (2022).

20. Taha, A. A. & Hanbury, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging* **15**, 1–28 (2015).